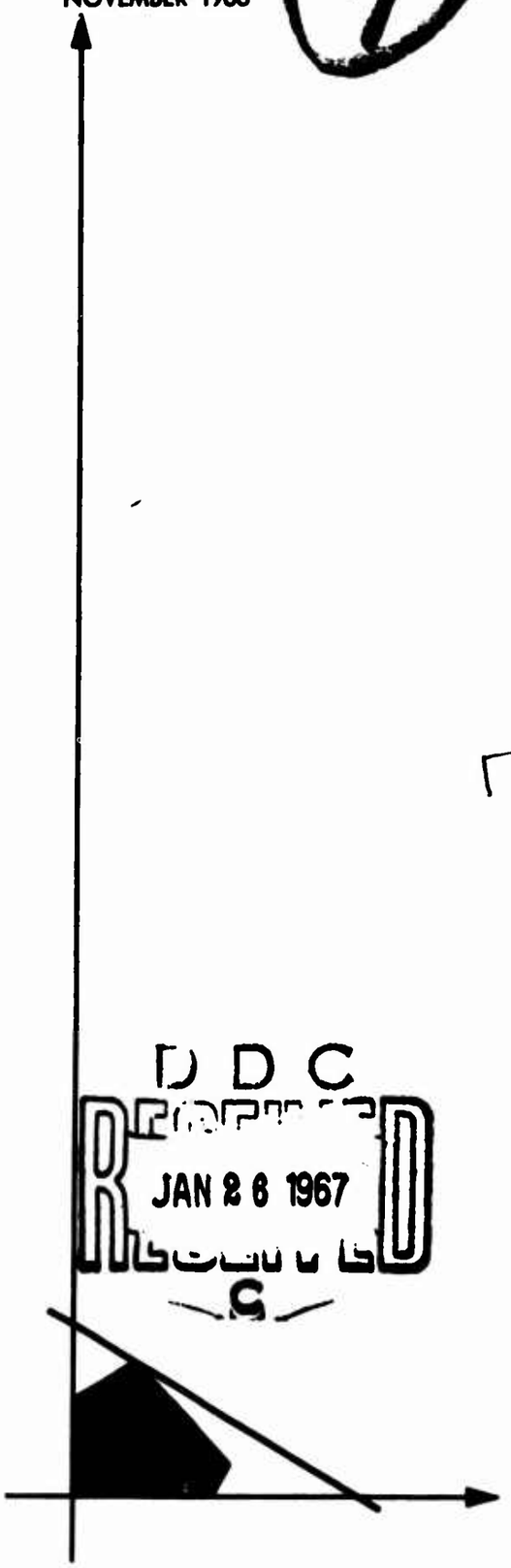
ORC 66-41
NOVEMBER 1966

A NOTE ON MARKOV-RENEWAL PROGRAMMING

by

William S. Jewell



OPERATIONS RESEARCH CENTER

COLLEGE OF ENGINEERING

UNIVERSITY OF CALIFORNIA - BERKELEY

A NOTE ON MARKOV-RENEWAL PROGRAMMING

by

William S. Jewell
Operations Research Center
University of California, Berkeley

November 1966 ORC 66-41

This research has been partially supported by the Army Research Office (Durham) under Contract No.: DA-31-124-ARO-D-331, and the National Science Foundation Grant GP-4593 with the University of California.

A NOTE ON MARKOV-RENEWAL PROGRAMMING

by

William S. Jewell

The various models of Markov-renewal programming were presented in [11], referred to thereafter as MRP I and MRP II; related models were presented independently by de Cani [2], Howard [10], and Schweitzer [15]. Since that time, new results have been obtained which clarify certain questions raised in these papers. Since these results are either "in the folklore," or are available only in scattered unpublished reports, it seemed worthwhile to gather them together in one article. We shall consider only the finite-state, finite alternative space, infinite-horizon, undiscounted and discounted models discussed in [11].

LINEAR PROGRAMMING FORMULATION AND RESULTS

It is well-known that Markov programs [9] can be represented as linear programs. The first such formulations are apparently due to Oliver [14], Manne [13], D'Épenoux [3], de Ghellinck [7], Wolfe and Dantzig [17], and Derman [4]. The extension of these formulations to Markov-renewal programs is straightforward, and the resulting primal and dual programs for both the undiscounted and discounted infinite-horizon cases are given in matrix form by Howard [10]. Because some of the details in the interpretation of the dual programs are not given, we consider these formulations, using the notation of [11].

In the case of undiscounted rewards, the primal problem is:

$$\text{Minimize} \quad x_0$$

$$(1) \qquad \text{Subject to:} \quad \sum_j (\delta_{ij} - p_{ij}^z)x_j + v_i^z x_0 \geq \rho_i^z \qquad \begin{pmatrix} i = 1,2, \ldots, N \\ \forall z \end{pmatrix}$$

$$x_j \text{ unrestricted} \qquad (j = 0,1, \ldots, N)$$

where $z = z(i)$ varies over all alternatives available in that state. (Usually we set $x_N = 0$.) At optimality, x_0 equals the maximal value of the gain rate, g , and the x_i equals the corresponding relative value

$$x_i = v_i = w_i - w_N \qquad (i = 1,2, \ldots, N) \tag{2}$$

in the limiting form of the total reward over $[0,t]$:

$$\lim_{t\to\infty} v_i(t) - gt = w_i + 0(1) \tag{3}$$

(We assume an ergodic underlying Markov chain. See (2)(B.6)(B.7)(12) in MRP II and and (100)(104) of [10].)

The dual to (1), after dividing by ν_i^z , is:

$$\text{Maximize} \quad \sum_i \sum_z y_i^z \left(\frac{\rho_i^z}{\nu_i^z}\right)$$

$$\text{Subject to:} \quad \sum_z y_j^z = \sum_z \sum_i y_i^z p_{ij}^z \qquad (j = 1,2, \ldots, N) \tag{4}$$

$$\sum_z \sum_i y_i^z = 1$$

$$y_i^z \geq 0 \qquad \left(\begin{matrix} i = 1,2, \ldots, N \\ \forall z \end{matrix}\right)$$

Directly from the constraints, we see that the y_i^z have the interpretation of "mixing coefficients" for the various alternatives in state i times the probability of being in that state; that is, if a <u>pure policy</u> were used, $y_i^z = \pi_i$ for some $z = z^*(i)$, and equals zero for all other alternatives available in that state. Then, since $\left(\rho_i^z/\nu_i^z\right)$ is the rate at which the reward is earned when in state i , following alternative z , the maximand is just the average rate at which reward is earned, at an arbitrary transition of the process--as it should be.

In the case of discounted reward, the optimal policy simultaneously maximizes the total discounted reward, starting in state i, $v_i(\alpha)$, for all states i. Thus, one can take any arbitrary set of initial starting probabilities, a_i $(i = 1,2, \ldots, N)$, and formulate the primal as:

$$\text{Minimize} \quad \sum_j a_j x_j$$

$$\text{Subject to:} \quad \sum_j \left(\delta_{ij} - p_{ij}^z \tilde{f}_{ij}^z(\alpha)\right) x_j \geq \rho_i^z \qquad \begin{pmatrix} i = 1,2, \ldots, N \\ \forall z \end{pmatrix} \tag{5}$$

$$x_j \quad \text{unrestricted} \qquad (j = 1,2, \ldots, N)$$

with the optimal values of x_i being the maximal values of the $v_i(\alpha)$. (This is a slight generalization of (111) of [10].) Directly, the dual of (5) is:

$$\text{Maximize} \quad \sum_i \sum_z y_i^z \rho_i^z$$

$$\text{Subject to:} \quad \sum_z y_j^z = a_j + \sum_z \sum_i y_i^z p_{ij}^z \tilde{f}_{ij}(\alpha) \qquad (j = 1,2, \ldots, N) \tag{6}$$

$$y_i^z \geq 0 \qquad \begin{pmatrix} i = 1,2, \ldots, N \\ \forall z \end{pmatrix}$$

To see the correct interpretation of (6), define

$$M_{ij}(t;\alpha) = \text{mean discounted number of entries into state } j \text{ in } [0,t], \text{ starting in state } i. \tag{7}$$

From first principles, or the undiscounted result (C.6) of MRP II (where $M_{ij}(t)$ did not include the event at the origin):

$$M_{ij}(t;\alpha) = \delta_{ij} + \sum_k \int_0^t e^{-\alpha x} M_{kj}(t-x;\alpha)\, dQ_{ik}(x) \tag{8}$$

In the limit, clearly

$$M_{ij}^{\alpha} \stackrel{\text{def}}{=} \lim_{t\to\infty} M_{ij}(t;\alpha) < \infty, \quad (\alpha > 0)$$

and

$$(9) \qquad M^{\alpha}_{ij} = \delta_{ij} + \sum_k p_{ik}\tilde{f}_{ik}(\alpha)M^{\alpha}_{kj} = \delta_{ij} + \sum_k M^{\alpha}_{ik}p_{kj}\tilde{f}_{kj}(\alpha) \quad .$$

The transposition follows from the well-known commutativity of the transforms of the matrices $Q_{ij}(t)$ and $M_{ij}(t)$ (undiscounted) [12]. If (9) were multiplied by the a_i and summed, an expression for the mean discounted number of visits to state j, under the starting conditions of (5), would be obtained. Comparing this with (6), we see that the dual variables y^z_i are just the appropriate alternative mixing probabilities times the discounted number of visits to state i; from this, it follows that the dual maximand in (6) is the correct value of total discounted return. For recent work on programming models, see [5], [6], and [8].

By straightforward use of the theory of linear programming, and the special structure of the constraint matrix in [4] and [9], it follows that pure strategies are optimal for both problems, a result first noted by Wagner [16] for Markov programs. In fact, the "policy improvement routine" is nothing more than a special version of the (dual) simplex method, in which simultaneous changes of several basis vectors (a basis is a selection of pure alternatives) are possible at each iteration, a fact noted by Oliver [14] and de Ghellinck [7]; no Phase I is needed, since any pure strategy is (dual) feasible. Schweitzer [15] has shown that it is the convention:

> "If there is no improvement in the test quantity from the last cycle, retain the same alternative"

in the policy-improvement algorithm which avoids cycling in the simplex method when there is degeneracy (tie among pure policies). Thus the algorithm is always finite.

In our discussion about alternative test criteria for the undiscounted case (Appendix D, MRP II), both Schweitzer and the author missed the "pricing-in" criterion which arises naturally from the linear programming models. It follows directly from (1) and (5) above that policy improvement will always occur if the

following rules are adopted for the algorithms of Figure 1, MRP I, and Figure 2, MRP II:

1. (Discounted, Infinite-Horizon Case): For at least one state i, select a new alternative $z(i)$ for which

$$v_i < \rho_i^z + \sum_j p_{ij}^z \tilde{f}_{ij}^z(\alpha) v_j \quad , \tag{10}$$

using the current values of the discounted returns, v_i.

2. (Undiscounted, Infinite-Horizon Case): For at least one state i, select an alternative $z(i)$ for which

$$v_i + g v_i^z < \rho_i^z + \sum_j p_{ij}^z v_j \quad , \tag{11}$$

using the current values of the gain, g, and the relative values, v_i.

Thus, both criteria (D.1) and (D.2) of MRP II are merely different ways of ranking prospective candidates to enter the basis at the next iteration, and the question of relative efficiency becomes undecidable without analysis of the computational labor required and experimental tests. Similar remarks apply to the question of rate of convergence if several candidates are placed in the basis simultaneously.

To summarize, the policy improvement algorithm of Markov and Markov-renewal programming is <u>simultaneously</u> a dynamic programming algorithm and a simplex algorithm.

TIES AND ABSOLUTE VALUES OF THE BIAS TERMS

Blackwell [1] remarked that the relative values in Markov programming were insufficient to break ties among policies with the same gains and produced an explicit formula for the absolute values of the RHS of (3), in terms of the "fundamental matrix" of Markov chains. The formula established for the absolute

values in (B.6) of MRP II for Markov-renewal programs was:

$$w_i = \lim_{t\to\infty} v_i(t) - gt = \sum_j \omega_{ij}\rho_j - \sum_j \left(\eta_j/\mu_{jj}\right) \tag{12}$$

where the notation is the same as in MRP II, except for the new definition:

$$\omega_{ij} = \frac{\mu_{ij}^{(2)}}{2\left(\mu_{jj}\right)^2} - \frac{\mu_{ij}}{\mu_{jj}} + \delta_{ij} \tag{13}$$

(The above limit may be in a Cesàro sense.)

This formula, involving the first and second moments of the first-passage distribution, is too complicated for rapid computation of the w_i ; similar remarks apply to a reduction of (B.6) by Schweitzer to a form involving the fundamental matrix (Equation (5.112) of [15]), and to remarks by Fox [5].

However, from basic definitions, and (C.8) of MRP II, we have:

$$\sum_i \sum_j \pi_i p_{ij} \nu_{ij} \omega_{jk} = \frac{\nu^{(2)}}{2\mu_{kk}} \tag{14}$$

where $\nu^{(2)}$ is the second moment of an average transition interval. Then, from (12), we have the remarkable formula:

$$\sum_i \sum_j \pi_i p_{ij} \nu_{ij} w_j = \frac{g\nu^{(2)}}{2} - \sum_j \pi_j \eta_j \quad . \tag{15}$$

(This result was first obtained in [12].)

Thus, once the stationary probabilities π_i are known, the normalizing factor to change the relative values v_i to the absolute values w_i follows directly. If the inverse of the matrix used in the value-determination part of the algorithm is saved, then the π_i may be determined from the row of the inverse corresponding

to the variable x_0 .

Although the above remarks do not make tie-breaking a trivial procedure, they do indicate that at most one need only carry out the value-determination procedure for every tying policy.

REFERENCES

[1] Blackwell, D., "Discrete Dynamic Programming," Ann.Math.Stat., Vol. 33, pp.719-726, (1962).

[2] de Cani, J. S., "A Dynamic Programming Algorithm for Embedded Markov Chains when the Planning Horizon is at Infinity," Management Science, Vol. 10, No. 4, pp.716-733, (1964).

[3] d'Épenoux, F., "Sur un problème de Production et de Stockage dans l'Aléatoire," Revue Française de Recherche Opérationelle, No. 14, pp.3-16 (1960). Translated and redrafted as "A Probabilistic Production and Inventory Problem," Management Science, Vol. 10, No. 1, pp.98-108, (1963).

[4] Derman, C., "On Sequential Decisions and Markov Chains," Management Science, Vol. 9, pp.16-24, (1962).

[5] Fox, B., "Markov Renewal Programming by Linear Fractional Programming," to appear in J.Soc.Ind.Appl.Math.

[6] _____, and E. V. Denardo, "Multichain Markov Renewal Programs," unpublished research memorandum, RAND Corporation, Santa Monica, (November 1966).

[7] de Ghellinck, G., "Les Problèmes de Décisions Séquentielles," Cahiers du Centre d'Etude de Recherche Opérationelle, Bruxelles, Vol. 2, No. 2, (1960).

[8] _____, and G. D. Eppen, "Linear Programming Solutions for Separable Markovian Decision Problems," to appear in Management Science.

[9] Howard, R. A., DYNAMIC PROGRAMMING AND MARKOV PROCESSES, Technology Press and Wiley Press, New York, (1960).

[10] _____, "Semi-Markovian Control Systems," Technical Report No. 3, Research in the Control of Complex Systems, Operations Research Center, Massachusetts Institute of Technology, (December 1963).

[11] Jewell, W. S., "Markov-Renewal Programming. I: Formulation, Finite Return Models," and "Markov-Renewal Programming. II: Infinite Return Models, Example," Operations Research, Vol. 11, No. 6, (1963), pp.938-948 and pp.949-971.

[12] _____, "Limiting Covariance in Markov-Renewal Processes," Research Report 64-16, O.R. Center, University of California, Berkeley, (July 1964).

[13] Manne, A. S., "Linear Programming and Sequential Decisions," Manag. Science, Vol. 6, pp.259-267, (1960).

[14] Oliver, R. M., "A Linear Programming Formulation of Some Markov Decision Processes," presented at a meeting of the Institute of Management Sciences and Operations Research Society of America, Monterey, California, (April 1960).

[15] Schweitzer, P. J., "Perturbation Theory and Markovian Decision Processes," Technical Report No. 15, Research in the Control of Complex Systems, Operations Research Center, Massachusetts Institute of Technology, (June 1965).

[16] Wagner, H. M., "On the Optimality of Pure Strategies," Management Science, Vol. 6, No. 3, pp.268-269, (1960).

[17] Wolfe, P., and G. B. Dantzig, "Linear Programming in a Markov Chain," Opns.Res., Vol. 10, pp.702-710, (1962).